Abnormal event detection in crowded scenes using sparse representation

Cong, Yang; Yuan, Junsong; Liu, Ji

2013

Cong, Y., Yuan, J., & Liu, J. (2013). Abnormal event detection in crowded scenes using sparse representation. Pattern recognition, 46(7), 1851-1864.

https://hdl.handle.net/10356/107417

https://doi.org/10.1016/j.patcog.2012.11.021

Abnormal Event Detection in Crowded Scenes using Sparse Representation

Yang Cong[†,‡], Junsong Yuan[‡] and Ji Liu[§]

[†] *State Key Laboratory of Robotics, Shenyang Institute of Automation, Chinese Academy of Sciences, China*

[‡] *Department of EEE, Nanyang Technological University, Singapore*

[§] *Department of Computer Sciences, University of Wisconsin-Madison, USA*

E-mail: congyang81@gmail.com, jsyuan@ntu.edu.sg, ji-liu@cs.wisc.edu [1]

Abstract

We propose to detect abnormal events via a sparse reconstruction over the normal bases. Given a collection of normal training examples, e.g., an image sequence or a collection of local spatio-temporal patches, we propose the sparse reconstruction cost (SRC) over the normal dictionary to measure the normalness of the testing sample. By introducing the prior weight of each basis during sparse reconstruction, the proposed SRC is more robust compared to other outlier detection criteria. To condense the over-completed normal bases into a compact dictionary, a novel dictionary selection method with group sparsity constraint is designed, which can be solved by standard convex optimization. Observing that the group sparsity also implies a low rank structure, we reformulate the problem using matrix decomposition, which can handle large scale training samples by reducing the memory requirement at each iteration from $O(k^2)$ to $O(k)$ where k is the number of samples. We use the column wise coordinate descent to solve the matrix decomposition represented formulation, which empirically leads to a similar solution to the group sparsity formulation. By designing different types of spatio-temporal basis, our method can detect both local and global abnormal events. Meanwhile, as it does not rely on object detection and tracking, it can be applied to crowded video scenes. By updating the dictionary incrementally, our

[1]This work was supported in part by the Nanyang Assistant Professorship (M4080134), JSPS-NTU joint project (M4080882), Natural Science Foundation of China (61105013), and National Science and Technology Pillar Program (2012BAI14B03). Part of this work was done when Yang Cong was a research fellow at NTU.

method can be easily extended to online event detection. Experiments on three benchmark datasets and the comparison to the state-of-the-art methods validate the advantages of our method.

Keywords:
sparse representation, abnormal event, crowd analysis, video surveillance

---

1. Introduction

Anomaly detection, also named as outlier detection, refers to detecting patterns in a given data set that do not conform to an established normal behavior, which is applicable in a variety of applications, such as intrusion detection, fraud detection, fault detection, system health monitoring, event detection in sensor networks, and detecting eco-system disturbances. The Oxford English Dictionary defines *abnormal* as:

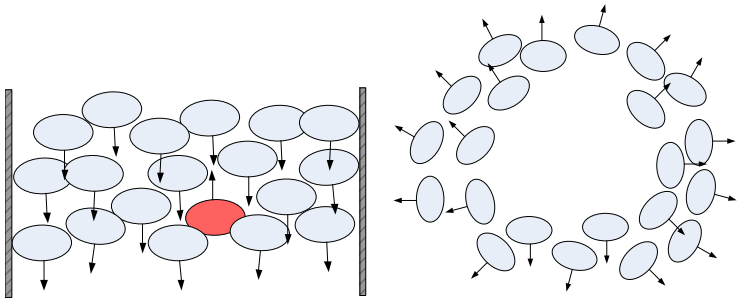
> *deviating from the ordinary type, especially in a way that is undesirable or prejudicial; contrary to the normal rule or system; unusual, irregular, aberrant*

We focus on the the detection of abnormal events in crowded scenes. According to the definition above, the abnormal events can be identified as irregular events from normal ones. Depending on the scale of interests, previous work in abnormal video event detection, such as [39, 3, 1, 23, 27, 4, 7], can be categorized into two classes, as shown in Fig.1 (each ellipse stands for a moving pedestrian):

i. Local abnormal event (LAE): the behavior of an individual is different from its neighbors. As shown in Fig.1(a), the motion pattern of the red one is different from its neighbors, thus is a spatial abnormal event.

ii. Global abnormal event (GAE): the group behavior of the global scene is abnormal. Fig.1(b) shows an abnormal scene, where the pedestrians suddenly scattered due to an abnormal event, e.g. an explosion.

Since the intention of each specific application is different, there is no unified definition for both local abnormal events and global abnormal event detection. Let us clarify the abnormal event detection firstly. Given the training set $\mathrm{D} = \{\mathrm{x}_1, \mathrm{x}_2, \ldots, \mathrm{x}_N\}$, where N is the number of training samples; $\mathrm{x}_i \in \mathbb{R}^d$ is a training data (d is the feature dimension), it stands for a general object which can be a



(a) Local Abnormal Event (LAE) (b) Global Abnormal Event (GAE)

Figure 1: The illustration of local and global abnormal events: each ellipse stands for a moving pedestrian. (a) Local Abnormal Event (LAE): the behavior of the red pedestrian is different from its neighbors. (b) Global Abnormal Event (GAE): the group behavior is abnormal.

pixel, an image patch, mixture dynamic texture, motion context in our paper, etc. Suppose we have a test sample $\mathrm{y} \in \mathbb{R}^d$, abnormal event detection is to design a measurement/function to determine whether y is normal or not. That is

$$f : \mathrm{y} \mapsto \{normal, abnormal\}. \tag{1}$$

To achieve this, two key issues need to be properly addressed, *event representation* and *anomaly measurement*.

For **abnormal event representation**, binary features based on background model are adopted in [39, 3]. Some other methods consider the spatial-temporal information, such as Histogram of Optical Flow (HOF) [1], spatial-temporal gradient [17], social force model [27], chaotic invariant [34], mixtures of dynamic textures [23]. There are also saliency feature [14] and graph-based non-linear dimensionality reduction method [30]. Moreover, the co-occurrence matrix is often used to describe the spatial relationship.

For **anomaly measurement**, to address this one-class learning problem, most conventional algorithms [1, 17, 16, 27] intend to detect testing sample with lower probability as anomaly by fitting a probability model over the training data. There are several statistics models, such as Gaussian model, Gaussian Mixture Model (GMM) or Mixture Principle Component Analysis (MPPCA) [16], Hidden Markov Model (HMM) [17], Markov Random Field (MRF) [3] or spatio-temporal MRF [16], Latent Dirichlet Allocation (LDA) [34]. Normalization Cut is used in [39] to discriminate the abnormal clusters from normal clusters. The procedure is to first fit some of stochastic probability model as mentioned above using the

training data set D, and then calculate the posterior probability of y given the model:

$$f = \begin{cases} \textit{normal} & p(\mathrm{y}|\mathrm{D}) \geq \theta \\ \textit{abnormal} & p(\mathrm{y}|\mathrm{D}) < \theta, \end{cases} \tag{2}$$

where θ is the threshold.

1.1. Motivation and Contribution

High-dimensional feature is usually preferred to better represent the event. However, to fit a good probability model, the required number of training data increases exponentially approximate $O(d^2)$ with the feature dimension d, it is unrealistic to collect enough training data for density estimation in practice. Thus, for most state-of-the-art methods, there is an unsolved problem between event representation using high-dimensional feature and model complexity. For example, for our global abnormal detection, there are only 400 training samples with the dimension of 320. With such a limited number of training samples, it is difficult to even fit a Gaussian model robustly.

We notice that, sparse representation is suitable to represent high-dimensional samples using less training data. This motivates us to detect abnormal events via a sparse reconstruction from normal ones. Given an input test sample $\mathrm{y} \in \mathbb{R}^m$, we reconstruct it by a sparse linear combination of an over-complete normal (positive) bases set $\Phi = \mathbb{R}^{m \times D}$, where $m < D$, as in Eq.(3):

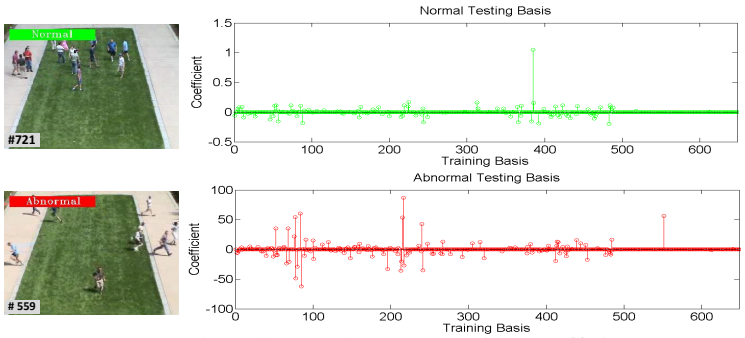
$$\mathrm{x}^* = \arg\min_{\mathrm{x}} \frac{1}{2}\|\mathrm{y} - \Phi\mathrm{x}\|_2^2 \ + \ \lambda\|\mathrm{x}\|_1, \tag{3}$$

where x^* is the reconstruction coefficients. As shown in Fig.2(a), a normal event (the up one) is likely to generate sparse reconstruction coefficients x^*, while an abnormal event (the bottom one) is dissimilar to any of the normal bases, thus generates a dense representation. To quantify the normalness, we propose a novel *sparse reconstruction cost* (SRC) based on the L_1 minimization, as

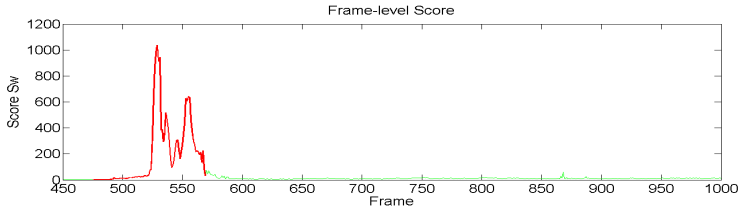
$$\mathrm{SRC} = \frac{1}{2}\|\mathrm{y} - \Phi\mathrm{x}^*\|_2^2 \ + \ \lambda\|\mathrm{x}^*\|_1. \tag{4}$$

As shown in Fig.2(b), for the frame-level abnormal event detection, the normal frame has a small reconstruction cost, while the abnormal frame usually generates a large reconstruction cost. Therefore, the SRC can be adopted as an anomaly measurement for such a one-class classification problem.

To handle both LAE and GAE, the definition of training basis y can be quite flexible, e.g., an image patch, a spatio-temporal video subvolume, or a normal



(a) Normal & Abnormal Reconstruction Coefficients



(b) SRC S_w in frame level

Figure 2: (a) Top left: the normal sample; top right: the sparse reconstruction coefficients; bottom left: the abnormal sample; bottom right: the dense reconstruction coefficients. (b) Frame-level Sparsity Reconstruction Cost (SRC): the red/green color corresponds to abnormal/normal frame, respectively. It shows that the S_w of abnormal frame is greater than normal ones, and we can identify abnormal events accordingly.

image frame. It thus provides a general way of representing different types of abnormal events. Moreover, we propose a new dictionary selection method to reduce the size of the basis of Φ for an efficient reconstruction of y. The weight of each new training sample is also learned to indicate its normalness, i.e., the occurrence frequency. These weights form a weight matrix W which serves as a prior term in the L_1 minimization.

We evaluate our method in three different abnormal event detection datasets, including the UMN dataset[31], the UCSD dataset[23], and the subway dataset[1]. The main contributions are as below:

i. For anomaly measurement, we propose a novel criterion, Sparse Reconstruction Cost (SRC), to detect abnormal event, which outperforms the existing criterion, e.g., Sparsity Concentration Index in [33]. The Weighted Orthogonal Matching Pursuit (WOMP) is also adopted to solve the weighted L_1 minimization in a more efficient way.

ii. To increase computational efficiency, a novel dictionary selection model based on group sparsity has been designed to generate a minimal size of bases set and prune noise training samples. Moreover, the lower rank constraint is considered to handle the large scale problem caused by large scale training samples.

iii. By using different types of bases, we provide a unified solution to detect both local and global abnormal events in crowded scene. Our method can also be extended to online event detection by an incremental self-update mechanism.

The rest of this paper is organized as follows: Section 2 gives the related work. Section 3 provides an overview of our algorithm. Section 4 presents the implementation details of our algorithm, including basis definition, dictionary selection, weighted L_1 minimization and self-update procedure. For dictionary selection, we compare the large scale version with the traditional one [7] in section 5. Then, section 6 reports our experimental results and comparisons with state-of-the-art methods to justify the performance of our algorithm. Finally, Section 7 concludes the paper.

2. Related Work

Much progresses in video surveillance have been achieved in recent years for some key areas, such as background model [29], object tracking [2], pedestrian detection [8], action recognition [36], crowd counting [6] and traffic monitoring [32]. Abnormal event detection, as a key application in video surveillance, has also provoked great interests. Depending on the specific scene, the abnormal event detection can be classified into those in **crowded scenes** and **uncrowded scenes**.

For **uncrowded scenario**, as the foreground objects can be extracted easily from the background, binary features based on background model are usually adopted, such as Normalized Cut clustering by Zhong et al.[39] and 3D spatio-temporal foreground mask feature fused using Markov Random Field by Benezeth et al.[3]. Due to the object template can be initialized in the uncrowded scene, there are also some trajectory-based approaches by tracking the objects, such as [32], [12], [32], [28] and [15]. They use frame-difference for object localization and then generate the object trajectories by tracking. These methods can obtain satisfied results on traffic monitoring, however, may fail in the crowded scene, since they cannot get a good object trajectories.

For **crowded scenes**, as there are so many objects or events occurring simultaneously in the clutter background, e.g. the subway station, it is difficult to separate each of objects or events and represent the overall object or event in global view. Therefore, most of the state-of-the-art methods use the local features for abnormal event representation, by considering the spatio-temporal information and extracting motion or gray-level sift-like features from local 2D patches or local 3D bricks, such as Histogram of optical flow, 3D gradient. Next the co-occurrence matrices are often chosen to describe the context information. For example, Adam et al.[1] use histograms to measure the probability of optical flow in local patch. Kratz et al. [17] extract spatio-temporal gradient to fit Gaussian model of each 3D brick, and then use HMM to detect abnormal events in densely crowded subway. Andrade et al. [10] use unsupervised feature extraction to encode normal crowd behaviour. The saliency features are extracted and associated using a Bayesian model to detect surprising (abnormal) events in video [14]. Kim et al.[16] model local optical flow with MPPCA and enforce consistency by Markov Random Field. In [30], a graph-based non-linear dimensionality reduction method using motion cues is applied for abnormality detection. Mahadevan et al.[23] model the normal crowd behavior by mixtures of dynamic textures. Mehran et al.[27] present a new way to formulate the abnormal crowd behavior by adopting the social force model [9, 35]. They first extract particle advection based on optical flow, then compute the social force and combine with a Latent Dirichlet Allocation (LDA) model for anomaly detection; however, their algorithm can just detect the global behavior in full image scale and cannot localize the sub-part abnormal region. In [34], they define a chaotic invariant to describe the event. Another interesting work is about irregularities detection by Boiman and Irani [4, 5], in which they extract 3D bricks as the descriptor and use dynamic programming as inference algorithm to detect the anomaly. Since this method searches the current feature from all the features in the past, it is time-consuming.

On the other hand, researchers have revealed that many neurons are selective for a variety of specific stimuli, e.g. color, texture, primitive, and this phenomenon broadly exists in both low-level and mid-level human vision [26, 25]. Therefore, sparse representation [26, 25, 18] is generated accordingly, which calls for modeling data vectors as a linear combination of a few elements from an overcomplete dictionary. Depending on the sparse reconstruction coefficients, sparse representation has also been used for many matching and classification applications in computer vision domain, such as object tracking [24], object or face recognition [22], image inpainting [20]. In comparison with conventional sparse representation, where the bases in dictionary are selected manually or generated by a dic-

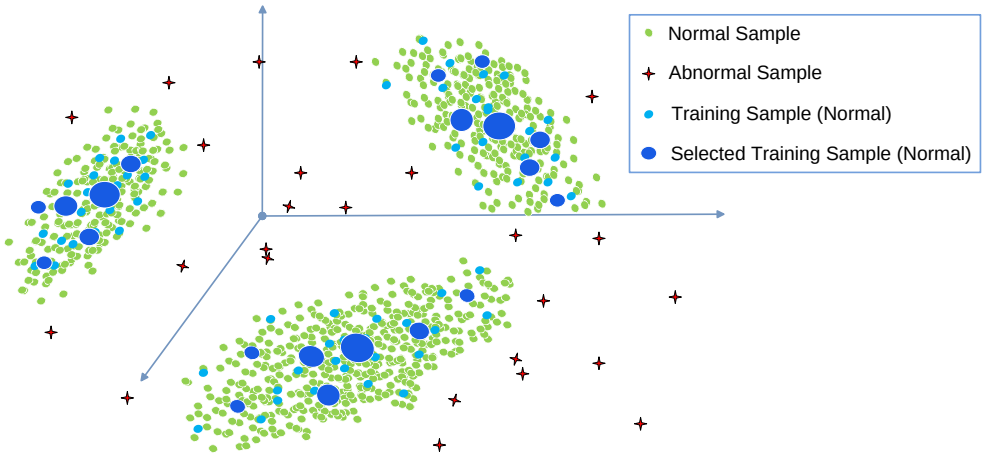


Figure 3: The illustration of our algorithm. Each point stands for a high dimensional feature point. The green or red point indicates the normal or abnormal testing sample, respectively. As most events are normal, the green points are dense and red points are sparse. For initialization the dictionary, some redundant light blue points are given as training features; after dictionary selection, an optimal subset of representatives (dark blue point) are selected as basis to constitute the normal dictionary, where its size indicates the weight: the larger, the more normal. Then, the abnormal event detection is to measure the sparsity reconstruction cost (SRC) of a testing sample (green and red points) over the normal dictionary (dark blue points).

tionary learning model, we propose a large scale dictionary selection model using low rank constraint, which can retain the original property of the data. Next, we propose a unified solution for abnormal event detection using sparse reconstruction cost (SRC) [7]. A similar work in [38] also applies sparse representation for abnormal event detection. However, it does not address the large-scale dictionary selection problem, and can not handle both LAE and GAE simultaneously as well.

3. Overview of Our Method

To detect both LAE and GAE, we propose a general solution using sparse representation, as illustrated in Fig.3. The flowchart of our algorithm is shown in Fig.4.

For training, only normal videos are required. To detect abnormal events from normal training samples, we collect the feature from training video frames to generate the normal feature pool B, where each sample in B is normal feature. Different features are designed for LAE or GAE (Sec.4.1). As the normal feature pool B is redundant and contains noisy features, an optimal subset B' with minimal size,

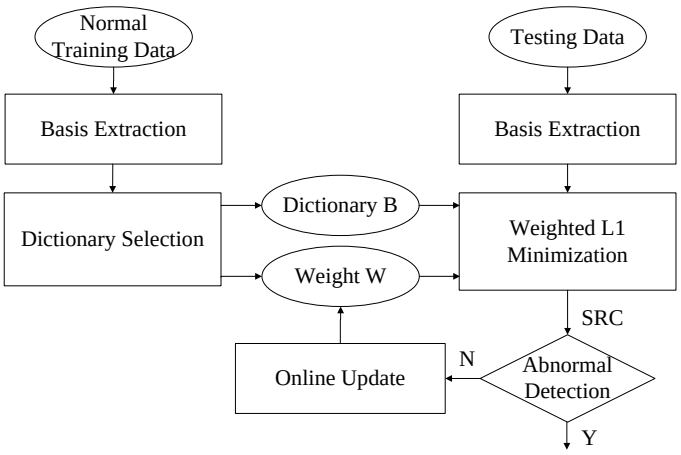


Figure 4: The flowchart of our proposed algorithm.

is selected from B as training dictionary (we call each feature of the selected dictionary as basis), and the weight of each basis of the dictionary is also initialized (Sec.4.2).

For testing, we also extract the same feature as in training, then each testing sample $\mathbf{y}$ can be a sparse linear combination of the training dictionary by weighted L_1 minimization, and whether $\mathbf{y}$ is to normal or not (e.g. the green/red point in Fig.3) is determined by the linear reconstruction cost (SRC) (Sec.4.3), i.e., normal feature can be efficiently spare represented by training dictionary with lower cost, on the contrary, the abnormal bases will be constructed with greater cost or even cannot be constructed, as shown in Fig.2. Moreover, our system can also self-update incrementally, which will be explained in Sec.4.4. The Algorithm is shown in Alg.2.

4. Implementation of Our Method

4.1. Multi-scale HOF and Basis Definition

We propose a new feature descriptor called Multi-scale histogram of optical flow (MHOF), and for event representation, all the types of bases, are concatenated by MHOF with various spatial or temporal structures. After estimating the motion field by optical flow [19], we partition the image into a few basic units, i.e. 2D image patches or spatio-temporal 3D bricks, then extract MHOF from each unit. For each pixel (x, y) of the unit, we quantize it into the MHOF as Eq.(5).

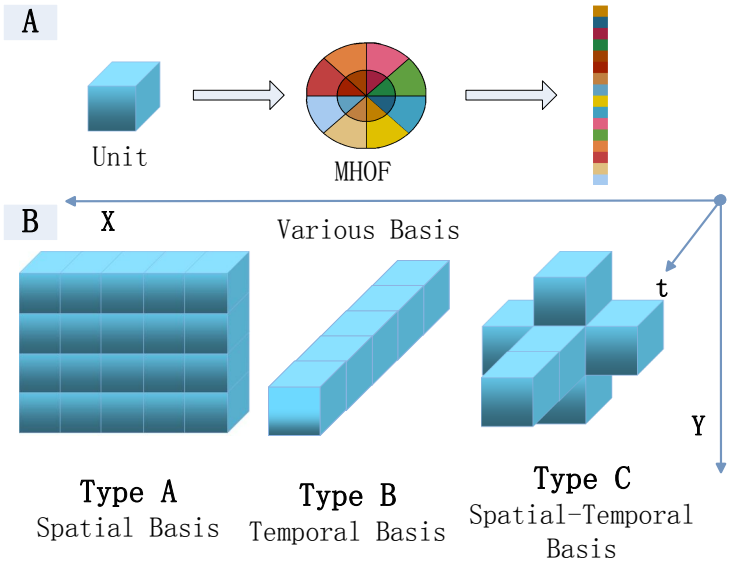


Figure 5: **(A)** The Multi-scale HOF is extracted from a basic unit (2D image patch or 3D brick) with 16 bins. **(B)** The selection of flexible spatio-temporal basis for sparse representation, such as type A, B and C, described by a concatenation of MHOF from the basic units. For GAE, we can use type A; for LAE, we can use type B or C.

In our implementation, the MHOF has $K = 16$ bins as shown in Fig.5: the first $p = 8$ bins denote 8 directions with motion energy $r < \tau$ in the inner layer; the next $p = 8$ bins correspond to $r \geq \tau$ ($\tau = 1$ in this paper) in the outer layer.

$$h(x,y) = \begin{cases} \text{round}(\frac{\text{p}\theta(\text{x,y})}{2\pi}) \bmod \text{p} & r(x,y) < \tau \\ \text{round}(\frac{\text{p}\theta(\text{x,y})}{2\pi}) \bmod \text{p} + \text{p} & r(x,y) \geq \tau \end{cases} \tag{5}$$

where $r(x,y)$ and $\theta(x,y)$ are the motion energy and motion direction of motion vector at (x,y) respectively. Therefore, our MHOF not only describes the motion information as traditional HOF, but also preserves the spatial contextual information. Actually depending on the specific applications, we can define much more scales MHOF, but for us, two scales are enough.

To handle different abnormal events, LAE or GAE, we propose several type of bases with different spatio-temporal structures, whose representations by the normalized MHOF is illustrated in Fig.5. For GAE, we select the spatial bases that can cover the whole frame. For LAE, we extract temporal or spatio-temporal bases that contain spatio-temporal contextual information, like the 3D Markov random field [16], and spatial topology structure can replace co-occurrance matrix. In general, our design of the local and global features is very flexible and other alternatives are certainly possible. Moreover, several features can be concatenated to build a more advanced description.

4.2. Large Scale Dictionary Selection using Sparsity Consistency

In this section, we address the problem of how to select the dictionary given an initial candidate feature pool as $\text{B} = [\text{b}_1, \text{b}_2, \cdots, \text{b}_k] \in \mathbb{R}^{m \times k}$, where each column vector $\text{b}_i \in \mathbb{R}^m$ denotes a normal feature. Our goal is to find an optimal subset to form the dictionary $\text{B}' = [\text{b}_{i_1}, \text{b}_{i_2}, \cdots, \text{b}_{i_n}] \in \mathbb{R}^{m \times n}$ where $i_1, i_2, \cdots, i_n \in \{1, 2, \cdots, k\}$, such that the set B can be well reconstructed by B' and the size of B' is as small as possible. A simple idea is to pick up candidates randomly or uniformly to build the dictionary. Apparently, this cannot make full use of all candidates in B. Also it is risky to miss important candidates or include the noisy ones, which will greatly affect the reconstruction. To solve this problem, we present a principled method to select the dictionary. Our idea is that we should select an optimal subset of B as the dictionary, such that the rest of the candidates can be well reconstructed using it. More formally, we formulate the problem as follows:

$$\min_{\text{X}} : \ \frac{1}{2}\|\text{B} - \text{BX}\|_F^2 + \lambda \|\text{X}\|_1, \tag{6}$$

where $\mathrm{X} \in \mathbb{R}^{k\times k}$; the Frobenius norm $\|\mathrm{X}\|_F$ is defined as $\|\mathrm{X}\|_F := (\sum_{i,j} \mathrm{X}_{ij}^2)^{\frac{1}{2}}$; and the L_1 norm is defined as $\|\mathrm{X}\|_1 := \sum_{i,j} |\mathrm{X}_{ij}|$. However, this tends to generate a solution of X close to an identity matrix I, which leads the first term of Eq.(6) to zero and is also very sparse. Thus, we need to require the consistency of the sparsity on the solution, i.e., the solution needs to contain some "0" rows, which means that the corresponding features in B are not selected to reconstruct any data samples.

Thus, in [7], we change the L_1 norm constraint in Eq.(6) into the $L_{2,1}$ norm, and propose the followed optimization problem to select the dictionary:

$$\min_{\mathrm{X}} : \frac{1}{2}\|\mathrm{B} - \mathrm{BX}\|_F^2 + \lambda \|\mathrm{X}\|_{2,1}, \tag{7}$$

where $\|\mathrm{X}\|_{2,1} := \sum_{i=1}^{k} \|\mathrm{X}_{i.}\|_2$, and $\mathrm{X}_{i.}$ denotes the i^{th} row of X. The regularization term enforces the group sparsity on the variable X and the optimal solution usually contains zero rows, i.e. the dictionary B' is constituted by selecting bases with $\|\mathrm{X}_{i.}\|_2 \neq 0$. The larger the value of λ is, the more zero rows X has. One can select the bases from the optimal X^* to build dictionary, i.e., the nonzero rows correspond to the selected basis. The $L_{2,1}$ norm is indeed a general version of the L_1 norm since if X is a vector, then $\|\mathrm{X}\|_{2,1} = \|\mathrm{X}\|_1$. In addition, $\|\mathrm{X}\|_{2,1}$ is equivalent to $\|\mathrm{x}\|_1$ by constructing a new vector $\mathrm{x} \in \mathbb{R}^k$ with $\mathrm{x}_i = \|\mathrm{X}_{i.}\|_2$. From this angle, it is not hard to understand that Eq.(6) leads to a sparse solution for X, i.e., X is sparse in terms of rows.

4.2.1. Improvement

This model in Eq.(7) looks pretty nice, but may lead to a memory problem when the number of samples is huge, since X requires k^2 units to save. When the value of k increases for the large scale problem, X cannot be loaded into the memory at one time, then Eq.(7) would be inefficient. In order to handle the large scale problem in practical applications, we decompose X as $\mathrm{X} = \alpha\beta^T$ where $\alpha \in \mathbb{R}^{k\times r}$ and $\beta \in \mathbb{R}^{k\times r}$ (r can be much smaller than k). Since the expected solution of X^* should contain many zero rows, which implies that it is also a low rank matrix, this decomposition $\mathrm{X} = \alpha\beta^T$ does not lose the generality. Typically, r can be given a number less than k, i.e. $r \ll k$. Thus, the memory cost in this decomposition is much less than k^2.

As we still desire a solution with multiple zero rows, the group sparsity constraint can be enforced on α. Apparently, the zero rows of X can be indicated by

those of α. Now we can reformulate Eq.(7) into a large scale version as follows:

$$\min_{\alpha,\beta}: \ \frac{1}{2}\|\mathrm{B}\alpha\beta^T - \mathrm{B}\|_F^2 + \lambda\|\alpha\|_{2,1}. \tag{8}$$

However, it is not enough because the optimal α^* would be infinitely close to 0 and the optimal β^* would be unbounded. To fix this problem, we only need one constraint on β such that β is bounded. Here, we can simply use the constraint $\|\beta\|_\infty \leq 1$ and formulate the completed version as follows:

$$\min_{\alpha,\beta}: \ \frac{1}{2}\|\mathrm{B}\alpha\beta^T - \mathrm{B}\|_F^2 + \lambda\|\alpha\|_{2,1} \ \text{ s.t.}: \ \|\beta\|_{\infty,\infty} \leq 1, \tag{9}$$

where $\|\beta\|_{\infty,\infty} := \max_{i,j}|\beta_{ij}|$. Note this problem is a nonconvex optimization problem, which can only guarantee a solution in the stationary point. The followed paragraph introduces the algorithm to solve this problem. Since there are two variables, we use the coordinate descent method to optimize α and β, iteratively, i.e., fixing α to optimize β and fixing β to optimize α, alternatively.

- **Optimize α:** While fixing β, we aim to solve the following subproblem:

$$\min_{\alpha}: \ F(\alpha) = \frac{1}{2}\|\mathrm{B}\alpha\beta^T - \mathrm{B}\|_F^2 + \lambda\|\alpha\|_{2,1}. \tag{10}$$

  This is a convex but nonsmooth optimization problem, as in our previous work [7]. Denote $f_0(\alpha)$ as the smooth part $\frac{1}{2}\|\mathrm{B}\alpha\beta^T - \mathrm{B}\|_F^2$. We employ the proximal method to solve it by the following updating procedure:

$$\begin{aligned}\alpha_{k+1} = &\arg\min_{\alpha}: p_{\alpha_k,L}(\alpha) \ := f_0(\alpha_k) + \langle \nabla f_0(\alpha_k), \alpha - \alpha_k\rangle \\ &+ \frac{L}{2}\|\alpha - \alpha_k\|^2 + \lambda\|\alpha\|_{2,1},\end{aligned} \tag{11}$$

  where L is the Lipschitz constant (or a larger number) and $\nabla f_0(\alpha_k)$ can be computed by $\mathrm{B}^T\mathrm{B}(\alpha_k\beta^T - \mathrm{I})\beta^T$. The closed form of α_{k+1} is given by $\mathscr{D}_{\frac{\lambda}{L}}(\alpha_k - \nabla f_0(\alpha_k)/L)$ due to the following theorem:

Theorem 1:

$$\arg\min_{\mathrm{X}} p_{\mathrm{Z,L}}(\mathrm{X}) = \mathscr{D}_{\frac{\lambda}{L}}(\mathrm{Z} - \frac{1}{L}\nabla f_0(\mathrm{Z})), \tag{12}$$

where $\mathscr{D}_\tau(.): \ \mathrm{M} \in \mathbb{R}^{k\times k} \mapsto \mathrm{N} \in \mathbb{R}^{k\times k}$

$$\mathrm{N}_{i.} = \begin{cases} 0, & \|\mathrm{M}_{i.}\| \leq \tau; \\ (1 - \tau/\|\mathrm{M}_{i.}\|)\mathrm{M}_{i.}, & \text{otherwise.} \end{cases} \tag{13}$$

Appendix.Appendix A gives the derivation to this theorem. Note that one can fix β and optimize α for multiple times, but the practical experiments indicate that one time is optimal.

- **Optimize β:** While fixing α to solve β, this subproblem is

$$\min_{\beta}: \ \frac{1}{2}\|\mathrm{B}\alpha\beta^T - \mathrm{B}\|_F^2 \ \text{ s.t.}: \ \|\beta\|_\infty \leq 1. \tag{14}$$

To optimize β we employ the idea "columnwise coordinate descent" in [21]: each column of β can be optimized simultaneously while fixing other columns:

$$\begin{aligned} &\min_{\beta_{.i}}: \ \frac{1}{2}\|\mathrm{B}\alpha\beta^T - \mathrm{B}\|_F^2 \equiv \frac{1}{2}\|(\mathrm{B}\alpha)_{.i}\beta_{.i}^T - (\mathrm{B} - \sum_{j\neq i}(\mathrm{B}\alpha)_{.j}\beta_{.j}^T)\|_F^2 \\ &\text{s.t.}: \ \|\beta_{.i}\|_\infty \leq 1. \end{aligned} \tag{15}$$

The optimal $\beta_{i.}^*$ can be computed by

$$\beta_{.i}^* = sgn(\mathrm{Z}) \odot \min(\mathrm{Z}, 1), \tag{16}$$

where $\mathrm{Z} = (\mathrm{B} - \sum_{j\neq i}(\mathrm{B}\alpha)_{.j}\beta_{.j}^T)^T(\mathrm{B}\alpha)_{.i}/\|(\mathrm{B}\alpha)_{.i}\|^2$. The operator "$\odot$" is defined by $(a \odot b)_i = a_i b_i$. Although we can update each column of β multiple times when optimize β, the practical results indicate that updating only once is usually enough.

We summarize the algorithm in Alg.1.

4.3. Anomaly measurement: weighted L_1 minimization and abnormal Detection

This section details how to determine a testing sample $\mathbf{y}$ to be normal or not. As we mentioned in the previous subsection, the features of a normal frame can be linearly constructed by only a few bases in the dictionary B' while an abnormal frame cannot. A natural idea is to pursue a sparse representation and then use the reconstruction cost to judge the testing sample. In order to advance the accuracy of prediction, two more factors are considered here:

- In practice, the deformation or any un-predicted situation may happen to the video. Motivated by [33], we extend the dictionary from B' to $\Phi = [\mathrm{B}', \mathrm{I}_{m\times m}] \in \mathbb{R}^{m\times D}$, and $D = n + m$.

Algorithm 1 Large Scale Dictionary Selection (LSDS)

Input: $\alpha_0 \in \mathbb{R}^{n\times r}$, $\beta_0 \in \mathbb{R}^{n\times r}$, $\lambda > 0$, K, L, r
Output: $\alpha_K \in \mathbb{R}^{n\times r}$, $\beta_K \in \mathbb{R}^{n\times r}$

1: **for** $k = 0, 1, 2, ..., K-1$ **do**
2: Update $\alpha_{k+1} = \mathscr{D}_{\frac{\lambda}{L}}(\alpha_k - \nabla f_0(\alpha_k)/L)$
3: **for** $i = 0, 1, 2, ..., n$ **do**
4: Update $(\beta_{k+1})_{.i} = sgn(\mathrm{Z}) \odot \min(\mathrm{Z}, 1)$

$$\mathrm{Z} = \mathrm{B} - \sum_{j=1}^{i-1} (\mathrm{B}\alpha_{k+1})_{.j} (\beta_{k+1})_{.j}^T - \sum_{j=i+1}^{n} (\alpha_{k+1})_{.j} (\beta_k)_{.j}^T$$

5: **end for**
6: **end for**

- If a basis in the dictionary appears frequently in the training dataset, then the cost to use it in the reconstruction should be lower, since it is a normal basis with high probability. Therefore, we design a weight matrix $\mathrm{W} = diag[w_1, w_2, ..., w_n, 1, ..., 1] \in \mathbb{R}^{D\times D}$ to capture this prior information. Each $w_i \in [0, 1]$ corresponds to the cost of the i^{th} feature. For the artificial feature set $\mathrm{I}_{m\times m}$ in our new dictionary Φ, the cost for each feature is set to 1. The way to dynamically update W will be introduced in the following section.

Now, we are ready to formulate this sparse reforestation problem:

$$\mathrm{x}^* = \arg\min_{\mathrm{x}} \frac{1}{2} \|\mathrm{y} - \Phi\mathrm{x}\|_2^2 \ + \ \lambda_1 \|\mathrm{W}\mathrm{x}\|_1, \tag{17}$$

where $\mathrm{x} = [\mathrm{x}_0, \mathrm{e}_0]^T$, $\mathrm{x}_0 \in \mathbb{R}^n$, and $\mathrm{e}_0 \in \mathbb{R}^m$. Given a testing sample y, we design a Sparsity Reconstruction Cost (SRC) using the minimal objective function value of Eq.(17) to detect its abnormality:

$$S_w = \frac{1}{2} \|\mathrm{y} - \Phi\mathrm{x}^*\|_2^2 \ + \ \lambda_1 \|\mathrm{W}\mathrm{x}^*\|_1. \tag{18}$$

A high SRC value implies a high reconstruction cost and a high probability of being an abnormal sample. In fact, the SRC function also can be equivalently mapped to the framework of Bayesian decision like in [13]. From a Bayesian view, the normal sample is the point with a higher probability, on the contrary the

Algorithm 2 Abnormal Event Detection Framework

Input: Training dictionary Φ, basis weight matrix W^0, sequential input testing sample $\mathrm{Y} \in [\mathrm{y}^1, \mathrm{y}^2, \cdots, \mathrm{y}^T]$

Output: W

1: **for** $t = 1, \cdots, \mathrm{T}$ **do**
2: Pursuit the coefficient x^* by L_1 minimization:
3: $\mathrm{x}^* = \arg\min_{\mathbf{x}} \frac{1}{2}\|\mathrm{y}^t - \Phi\mathrm{x}\|_2^2 + \|\mathrm{W}^{t-1}\mathrm{x}\|_1$
4: Calculate SRC function S_w^t by Eq.(18)
5: **if** y is normal **then**
6: Select top K bases coefficients of x^*
7: Update $\mathrm{W}^t \longleftarrow \mathrm{W}^{t-1}$
8: **end if**
9: **end for**

abnormal (outlier) sample is the point with a lower probability. We can estimate the normal sample by maximizing the posteriori as follows:

$$\begin{aligned}
\mathbf{x}^\star &= \arg\max_{\mathrm{x}} p(\mathrm{x}|\mathrm{y}, \Phi, \mathrm{W}) \\
&= \arg\max_{\mathrm{x}} p(\mathrm{y}|\mathrm{x}, \Phi, \mathrm{W}) p(\mathrm{x}|\Phi, \mathrm{W}) \\
&= \arg\max_{\mathrm{x}} p(\mathrm{y}|\mathrm{x}, \Phi) p(\mathbf{x}|\mathbf{W}) \\
&= \arg\min_{\mathrm{x}} -[\log p(\mathrm{y}|\mathrm{x}, \Phi) + \log p(\mathrm{x}|\mathrm{W})] \\
&= \arg\min_{\mathrm{x}} (\frac{1}{2}\|\mathrm{y} - \Phi\mathrm{x}\|_2^2 + \lambda_1 \|\mathbf{W}\mathbf{x}\|_{\mathbf{1}}),
\end{aligned} \tag{19}$$

where the first term is the likelihood $p(\mathrm{y}|\mathrm{x}, \Phi) \propto \exp(-\frac{1}{2}\|\mathrm{y} - \Phi\mathrm{x}\|_2^2)$, and the second term $p(\mathrm{x}; \mathrm{W}) \propto \exp(-\lambda_1 \|\mathrm{W}\mathrm{x}\|_1)$ is the prior distribution. This is consistent with our SRC function, as the abnormal samples correspond to smaller $p(\mathrm{y}|\mathrm{x}, \Phi)$, which results in greater SRC values.

4.3.1. Optimization

In our previous work [7], Eq.(17) can be solved by quadratic programming using the interior-point method, which uses conjugate gradients algorithm to compute the optimized direction. However, as solving Eq.(17) is time consuming for abnormal event detection, we need to a more efficient algorithm. To achieve this, the greedy algorithm for least squares regression in [37], called orthogonal matching pursuit (OMP) in signal processing community, is a good choice. Motivated

Algorithm 3 Weighted Orthogonal Matching Pursuit (WOMP)

Input: $\Phi \in \mathbb{R}^{m\times D}$, $y \in \mathbb{R}^m$, $\mathrm{W} \in \mathbb{R}^{D\times D}$, and $\varepsilon > 0$
Output: $\mathrm{x}^\star$, k, F

```
Normalize b̃_j = b_j / ||b_j||_2
K = 0, F = ∅ and x = 0
while (||Φx − y||_2 / ||y||_2 > ε) do
    k = k + 1
    i = arg max_i w_ii ||x̃_i^T (Φx − y)||
    Let F = {i} ∪ F
    Let x = (Φ_F^T Φ_F)^{-1} Φ_F^T y
end while
x* = x
```

by OMP [37], we design a Weighted Orthogonal Matching Pursuit (WOMP) model in our case. The improvement is that we change the second term of Eq.(17) by adding a weighted factor. The algorithm is shown in Alg.3.

Typically, we set $\varepsilon = 0.05$, which measures the reconstruction accuracy; F returns the support set; and $\mathrm{x}^\star$ is the pursuited parameter vector. Thus, we can use either weighted $L1$ minimization in [7] or the improved version of WOMP in Alg.3 to solve Eq.(17).

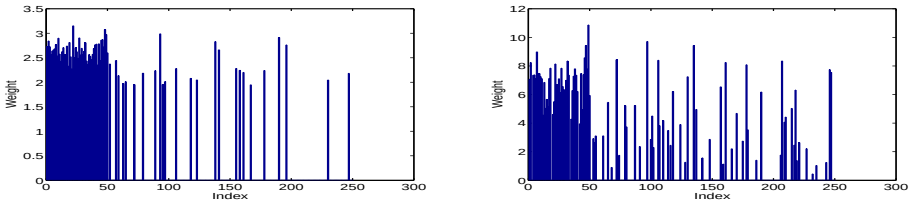
4.4. Update Weight and Dictionary

For the normal sample **y**, we selectively update weight W and dictionary Φ by choosing the top K bases with largest positive coefficient of $\mathrm{x}_0^* \in \mathbb{R}^n$, and we define the top k set as $S_k = [s_1, \cdots, s_k]$.

As we have mentioned above, the contribution of each basis to the L_1 minimization reconstruction is not identical. In order to measure such a contribution, we use W to assign each basis a weight, that is the basis with higher weight, should be used frequently and of course more similarity to normal event and vice verse. Define $W = [w_1, w_2, \cdots, w_K]$. We initialize W from matrix X of dictionary selection, that is

$$\beta_i^0 \;=\; \|\mathrm{X}_{i.}\|_2, \quad w_i^0 \;=\; 1 - \frac{\beta_i^0}{\|\beta^0\|_1}. \tag{20}$$

where β_i denotes the accumulate coefficients of each basis, and $w_i \in [0,1]$ (the smaller the value of w_i, the more like a normal sample it is). The top k basis in W



(a) The result of traditional model [7] (b) The result of our large scale model

Figure 6: The comparison of different dictionary selection models. The dimension of each basis is $m = 100$, $n_1 = 50$ and $n = 250$. The first $n_1 = 50$ samples are the basis as ground truth, the other $n_2 = 200$ samples are testing samples. (a) is the result of traditional model [7] and (b) is the result of our large scale dictionary model. We can see that most basis are selected from testing samples successfully for both (a) and (b).

	m	n_1	n	Convex Model [7]	**LSDS**
1	50	20	220	0.80	0.90
2	75	30	230	0.94	0.92
3	100	50	250	0.85	0.90

Table 1: The comparison of our proposed method with traditional model [7] for dictionary selection using synthesized data set.

can be updated as follows:

$$\beta_i^{t+1} = \beta_i^t + \mathrm{x}_i^*, \ \{i \in S_k\}, \quad w_i^{t+1} = 1 - \frac{\beta_i^{t+1}}{\|\beta^{t+1}\|_1}. \tag{21}$$

where S_k is the index set of the top k features in W.

5. Comparison Dictionary Selection model: LSDL Vs. Convex Model [7]

Based on the same dictionary selection model in Eq.(7), we have two optimization schemes, namely the large scale version in Eq.(9) and the traditional one in Eq.(7) [7]. The main difference between them is the memory cost, i.e. $\mathrm{X} = \alpha\beta$ relates to $2(k \times r), r \ll n$ and B relates to $k \times n$, which is crucial especially when n is large. If they have similar performance, i.e. they can select similar dictionary from the same testing samples, the traditional version can be replaced by the large scale version directly. Therefore, we compare our new large scale version with our previous work in [7] using synthesized data.

The experiment is setup in the following. Suppose the dimension of each sample as m, we first randomly generate n_1 basis, which can be considered as ground

truth; then we randomly linearly combine them to generate new n_2 samples; finally, we normalize them and have total of $n = n_1 + n_2$ samples. The accuracy is the number of bases selected in the proportion of n_1 ground truth. If the result of our large scale version is similar to [7], we can consider they have competitive performance. Moreover, in comparison with [7], the large scale version needs less memory cost and can also work well when n increases.

The simulation results are as shown in Fig.6 and some statistic results are provided in Tab.1. Thus, without considering the size of B, we can conclude that both of Eq.(7) and Eq.(9) have nearly the same performance.

6. Experiments

In this section, we systematically apply our proposed algorithm to several published data sets to justify the effectiveness. The UMN dataset [27] is used to test the Global Abnormal Event (GAE) detection; and the UCSD dataset [23, 11] and Subway datasets [1] are applied to Local Abnormal Event (LAE) detection. Moreover, we re-annotate Subway dataset in a bounding box level ground truth, where each box contains one abnormal event. For evaluation, three different level measurements are applied, which are Pixel-level, Frame-level and Event-level measurements.

6.1. Dataset

UMN dataset: The UMN dataset [27] consists of 3 different scenes of crowded escape events, and the total frame number is 7740 (1450, 4415 and 2145 for scenes $1-3$, respectively) with a 320×240 resolution. The normal events are pedestrians walking randomly on the square or in the mall, and the abnormal events are human spread running at the same time. There are total of 11 abnormal events in the whole video set.

UCSD dataset: The UCSD dataset [23, 11] includes two sub-datasets, Ped1 and Ped2. The crowd density varies from sparse to very crowded. The training sets are all normal events and contain only pedestrians. The abnormal events in testing set are either 1) the circulation of non pedestrian entities in the walkways, or 2) anomalous pedestrian motion patterns. Commonly occurring anomalies include bikes, skaters, small cars, and people walking across a walkway or in the grass that surrounds it. Due to Ped2 sub-dataset has no pixel-level ground truth, in this paper we mainly focus on Ped1. For Ped1, the training set includes 34 normal video clips and the testing set contains 36 video clips in which some of the frames have one or more anomalies presents (a subset of 10 clips in testing set are provided

with pixel-level binary masks to identify the regions containing abnormal events). For each clip, there are about 200 frames with the resolution 158×238, The total number of anomalies frames(≈ 3400) is a little bit smaller than that of normal frames (≈ 5000).

Subway dataset: The subway dataset is provided by Adam et al. [1], including two videos: "entrance gate" (1 hour 36 minutes long with 144249 frames) and "exit gate" (43 minutes long with 64900 frames). In our experiments, we resized the frames from 512×384 to 320×240. The abnormal events mainly include wrong direction events and no-payment events.

6.2. Evaluation Criterion

Three criteria in different levels are applied for evaluation, which are Pixel-level, Frame-level and Event-level.

- Pixel-level: To test localization accuracy, detections are compared to pixel-level ground truth masks, on a subset of ten clips. The procedure is similar to that described above. If at least 40% of the truly anomalous pixels are detected, the frame is considered detected correctly, and counted as a false positive otherwise.

- Frame-level: If a frame contains at least one abnormal pixel, it is considered as a detection. These detections are compared to the frame-level ground truth annotation of each frame. Note that this evaluation does not verify whether the detection coincides with the actual location of the anomaly. It is therefore possible for some portion true positive detections to be "lucky" co-occurrences of erroneous detections and abnormal events.

- Event-level: Usually, an abnormal event will last for several consecutive frames, if more than one frames are detected as abnormal and the position is localized exactly, it is considered as a detection. Note that this evaluation does not need all the abnormal frames are detected.

The Receiver Operating Characteristic (ROC) curve is used to measure the accuracy for multiple threshold values. The ROC is consisted of true positive rate (TPR) and false positive rate (FPR), of which TPR determines a classifier or a diagnostic test performance on classifying positive instances correctly among all positive samples available during the test, and FPR, on the other hand, defines how many incorrect positive results occur among all negative samples available during the test. These measures are given by the formulas in Eq.(22):

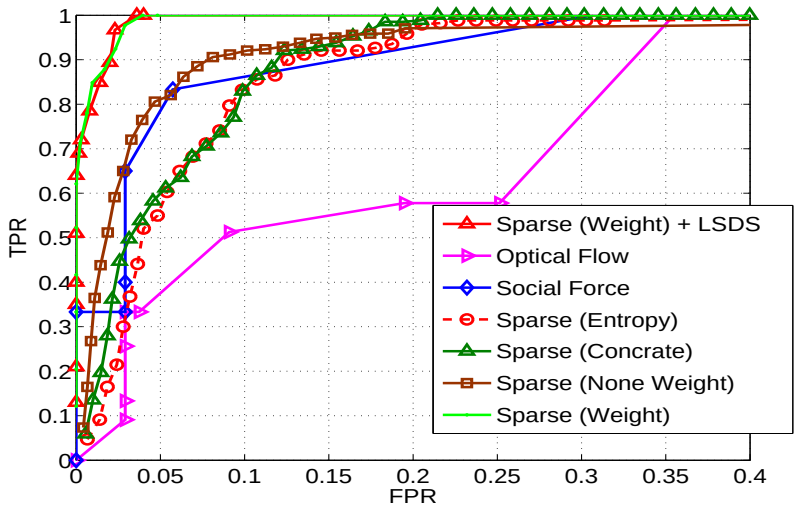


Figure 7: The ROCs for the detection of abnormal frames in the UMN dataset. We compare different evaluation measurements for abnormal event detection, namely weighted SRC with Large Scale Dictionary Selection model (LSDS), weighted SRC, SRC without Weight, sparse with concentration function and sparse with entropy measurement, and also the other two methods, social force [27] and optical flow [27]. Our method outperforms the others.

$$\begin{aligned} \text{TPR} &= \frac{\text{True positive}}{\text{True positive} + \text{False negative}} \\ \text{FPR} &= \frac{\text{False positive}}{\text{False positive} + \text{True negative}}, \end{aligned} \tag{22}$$

where True positive (TP) is the correctly labeled abnormal events; False negative (FN) is incorrectly labeled normal events; False positive (FP) is incorrectly labeled abnormal events; and True negative (TN) is correctly labeled abnormal events. For pixel-level and frame-level, we choice different thresholds and compute the TPR and FPR accordingly to generate the ROC curve.

6.3. Global Abnormal Event Detection

For the UMN dataset [27], we initialize the training dictionary from the first 400 frames of each scene, and leave the others for testing. The type A basis in

Method	AUC
Chaotic Invariants [34]	0.99
Social Force[27]	0.96
Optical flow [27]	0.84
1-NN	0.93
Sparse Scene1	0.995
Sparse Scene2	0.975
Sparse Scene3	0.964
Sparse+LSDS Scene1	**0.9955**
Sparse+LSDS Scene2	**0.971**
Sparse+LSDS Scene3	**0.974**

Table 2: The comparison of our proposed method with the state-of-the-art methods for detection of the abnormal events in the UMN dataset. We can see that our method with or without LSDS get similar results, but LSDS can be also used in the case that the size of training data is bigger.

Fig.5(B), i.e., spatial basis, is used here. We split each image into 4×5 blocks, and extract the MHOF from each block. We then concatenate them to build a basis with a dimension $m = 320$. Because the abnormal events cannot occur only in one frame, a temporal smooth is applied.

The results are shown in Fig.8. The normal/abnormal results are annotated as red/green color in the indicated bars respectively. In Fig.7, the ROC curves by frame-level measurement are shown to compare our SRC to three other measurements, which are

i. SRC with W as an identity matrix in Eq.(18), where $S = \frac{1}{2}\|\mathrm{y} - \Phi\mathrm{x}^*\|_2^2 + \lambda_1\|\mathrm{x}^*\|_1$.

ii. the entropy used as a metric by formulating the sparse coefficient as a probability distribution: $S_E = -\sum_i p_i \log p_i$, where $p(i) = |x(i)|/\|\mathrm{x}\|_1$. Thus sparse coefficients will lead to a small entropy value.

iii. concentration function similar to [33], $S_S = T_k(\mathrm{x})/\|\mathrm{x}\|_1$, where $T_k(\mathrm{x})$ is the sum of the k largest positive coefficients of x (the greater the $\mathrm{S_s}$, the more likely a normal testing sample).

Moreover, Tab.2 provides the quantitative comparisons to the state-of-the-art methods. The AUC of our method without using LSDS and using LSDS are similar, which are from 0.964 to 0.995 and from 0.971 to 0.9955, respectively; and both of them outperform [27] and are comparable to [34]. However, our method is a

more general solution, because it covers both LAE and GAE. Moreover, Nearest Neighbor (NN) method can also be used in high dimensional space by comparing the distances between the testing sample and each training samples. The AUC of NN is 0.93, which is lower than that of our method. This demonstrates the robustness of our sparse representation method over NN method.

6.4. Local Abnormal Event Detection

6.4.1. UCSD Ped1 Dataset

For the UCSD Ped1 dataset, we split each image into local patches of size 7×7 with 4 pixel overlaps. For event representation, we select type C basis in Fig.5 for incorporating both local spatial and temporal information, with the dimension $m = 7 \times 16 = 102$. From each localization, we estimate a dictionary and use it to determine whether a testing sample is normal or not. A spatio-temporal smooth is adopted here for eliminating noise, which can be seen as a simplified version of spatio-temporal Markov Random Filed [16].

Some testing results are shown in Fig.9, where both our approach with and without LSDS get satisfied results and outperform the state-of-the-art. Our approach can detect abnormal events such as bikers, skaters, small cars, etc. In Fig.10, we compare our method with MDT, Social force and MPPCA in [23] by using pixel-level and frame-level measurements defined in [23]. It shows that our ROC curve is better than others, and our approach with or without LSDS get similar results. In Tab.3, the performance is evaluated using different criteria: for the Equal Error Rate (EER), our method using LSDS is 20%, which is higher than our method without using LSDS but lower than other methods 25%[23]; for Rate of Detection (RD), ours using LSDS is the same as our previous version 46% and higher than the state-of-the-art methods 45% [23]; and for Area Under Curve (AUC), ours make a bit improvement from 46.1% to 48.7%, and both of them outperforms other methods, e.g 44.1%[23]. Therefore, it demonstrates that our algorithm outperforms the state-of-the-art methods.

6.4.2. Subway Dataset

For the subway dataset [1], we resized the frames from 512×384 to 320×240 and divided the new frames into 15×15 local patches with 6 pixel overlaps. For event representation, the type B basis in Fig.5 is adopted with the dimension of $m = 16 \times 5 = 80$. The first 10 minute video is collected for estimating an optimal dictionary. The patch-level ROC curves of both two data sets are presented in Fig.13, where the positive detection and false positive correspond to each individual patch, and the AUCs are about 80% and 83%.

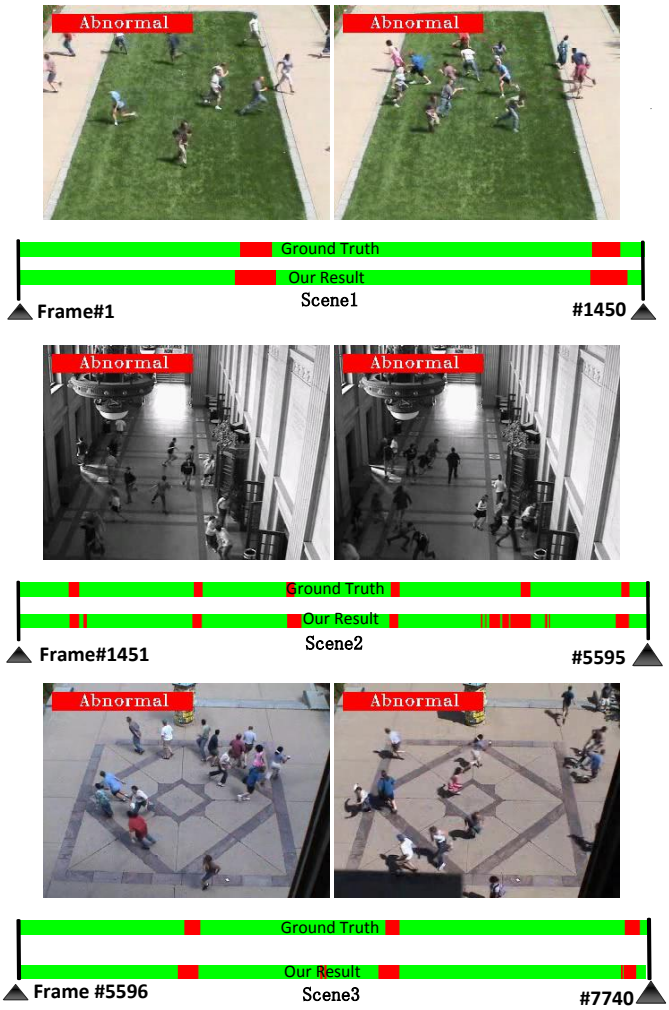


Figure 8: The qualitative results of the global abnormal event detection using our method with LSDS for three sample videos from UMN dataset, which is similar to Fig.3 in [7]. The top row represents the result for a video in the dataset. The ground truth bar and the detection result bar represent the labels of each frame for that video, and green color denotes the normal frames and red corresponds to abnormal frames.

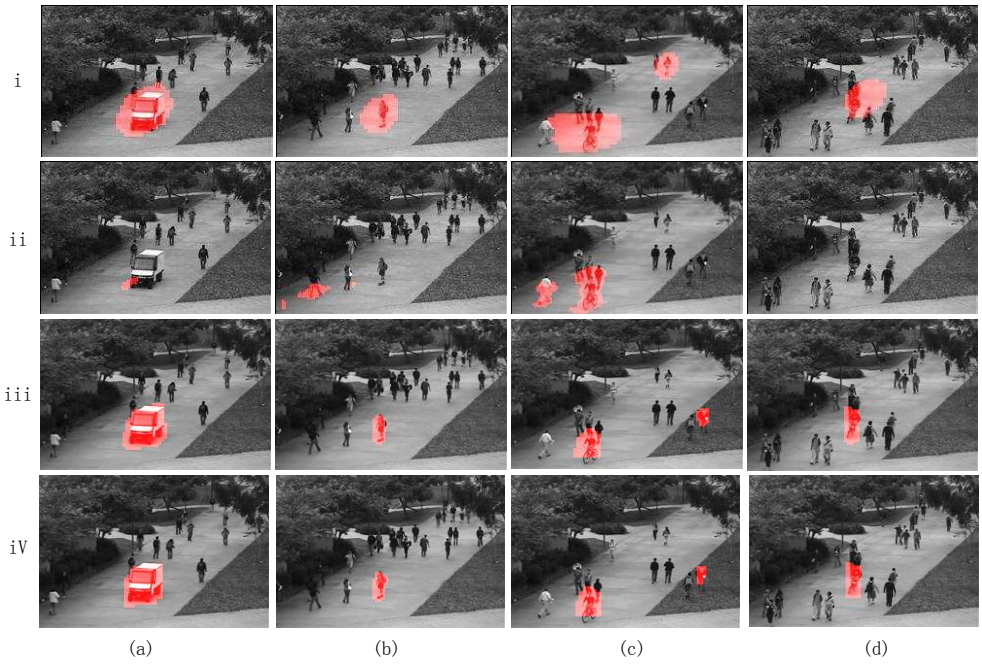


Figure 9: Examples of abnormal detections using (i) the MDT approach [23], (ii) the SF-MPPCA approach [23], (iii) our approach without Large Scale Dictionary Selection (LSDS) and (iv) our approach with LSDS, where our approach with or without LSDS get similar results. For MDT, its results are not accurate, which contain many background regions; and for SF-MPPCA, it completely misses the skater in (b), the person running in (c) and the biker in (d); moreover, both MDT and SF-MPPCA miss the person walking on the grass in (c). In contrast, our approach using sparse representation can outperform the state-of-the-art methods and obtain satisfactory results,

The detected results are shown in Fig.11. In additional to wrong direction events, the no-payment events are also detected, which are very similar to the normal "checking in" action. The event-level evaluation is shown in Tab.4, which are divided into two parts depending on different groundtruth definitions. Only our method can detect all the wrong direction events accurately. Moreover, in contrast to [1], our approach can also keep a higher accuracy for no-payment events, because the designed temporal basis contains the temporal causality context. For the measurement of false alarm, our method is also the lowest one.

For these three dataset of GAE and LAE, we find that our improved version using LSDS gets similar result as our previous one [7], this is because both of them select the most efficient bases to construct the dictionary and use them for sparse reconstruction. However, our LSDS can handle large scale training data,

	EER	RD	AUC
SF [23]	31%	21%	17.9%
MPPCA [23]	40%	18%	20.5%
SF-MPPCA [23]	32%	18%	21.3%
MDT [23]	25%	45%	44.1%
Adam[1]	38%	24%	13.3%
Sparse	**19**%	**46**%	**46.1**%
Sparse+LSDS	**20**%	**46**%	**48.7**%

Table 3: The statistical result of UCSD Ped1 dataset. Quantitative comparison of our method with [23]: EEE is equal error rate, RD is rate of detection, and AUC is the area under ROC.

	Wrong Direction	No-Pay	Total	False Alarm
Ground truth [1]	21/9	10/-	31/9	-/-
Adam[1]	17/9	-/-	17/9	4/2
Ours	**21/9**	**6**/-	**27/9**	**4/0**
Ground truth [16]	26/9	13/3	39/12	-/-
Kim [16]	24/9	8/3	32/12	6/3
Zhao [38]	25/9	9/3	34/12	5/2

Table 4: Comparison of accuracy for both subway videos. The first number in the slash (/) denotes the entrance gate result; the second is for the exit gate result. Due to different groundtruth annotations [1, 16], the table is classified into two parts. Nevertheless, our method is more accurate and has low false alarms rate than the state-of-the-art methods.

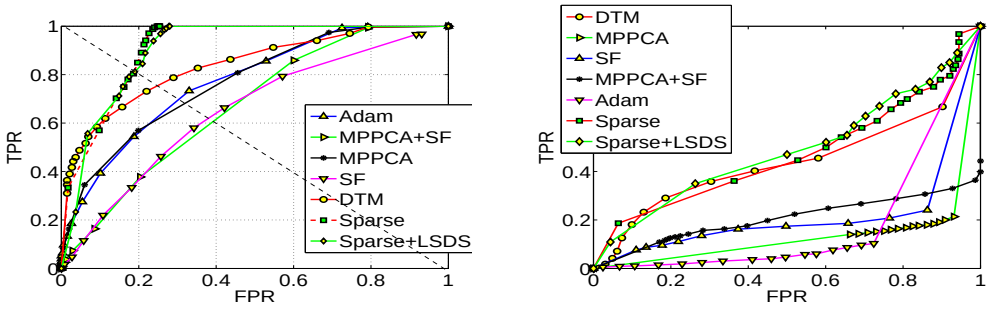


Figure 10: The result of UCSD Ped1 dataset. (a) Frame-level ROC for Ped1 Dataset, (b) Pixel-level ROC for Ped1 Dataset.

which is crucial in practical applications. All experiments are run on the computer with 2GB RAM and 2.6GHz CPU. The average computation time is 0.8 s/frames for GAE, 3.8 s/frame for UCSD dataset, and 4.6 s/frame for the Subway dataset.

6.5. Comparison: $L_{2,1}$ Norm Vs. Frobenius Norm

Some readers may ask why we use group sparsity, and whether sparsity is really effective or not. To answer these questions, we define a similar dictionary selection model using Frobenius norm to compare with our dictionary selection model using group sparsity, i.e. $L_{2,1}$, as below:

$$F_s = \arg\min_{\mathrm{X}} \frac{1}{2}\|\mathrm{B}-\mathrm{BX}\|_F^2 + \lambda_2\|\mathrm{X}\|_F^2, \tag{23}$$

where $\mathrm{X} \in \mathbb{R}^{k\times k}$ and $\mathrm{B} \in \mathbb{R}^{m\times k}$. To pursuit X, we can get a close-form solution as Eq.(24), which can be proved in Appendix.Appendix B:

$$\mathrm{X} = (\mathrm{B}^T\mathrm{B} + \lambda_2\mathrm{I})^{-1}\mathrm{B}^T\mathrm{B}. \tag{24}$$

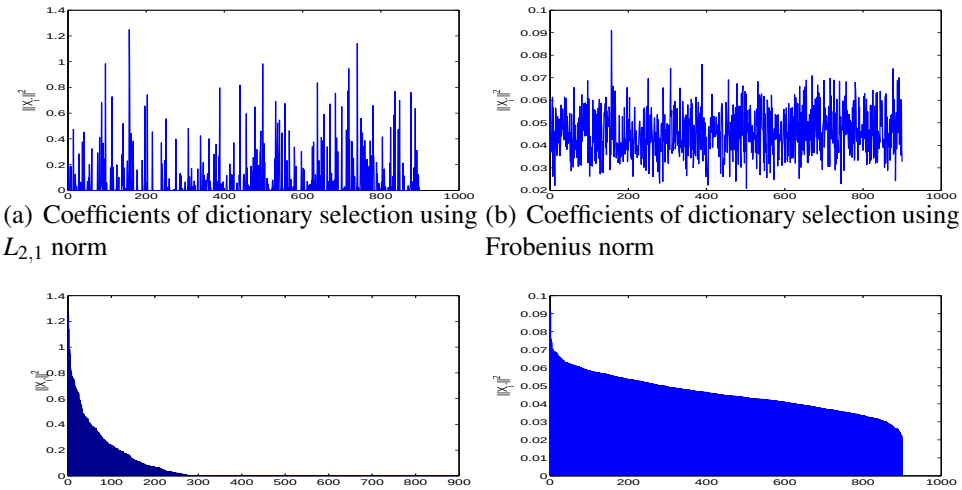
Now, let us compare our dictionary selection model using group sparsity with the Frobenius norm version in Eq.(23) using synthesized data. The original feature set B with each column as an independently feature is collected from three gaussian models, where the mean and covariance matrix of each gaussian model is randomly generated. Then we randomly sample each gaussian model to generate B. In detail here, each feature dimension is $m = 50$, and we randomly sample 300 features from each gaussian model, so that we totally have 900 candidate features. Then we set $\lambda = \lambda_2$ and give them different values to compare the results. A demo

Figure 11: Example abnormal event detected by our algorithm. The top row and bottom row are from exit and entrance video set, respectively, and red masks contained into the yellow rectangle indicate where the abnormal is detected, including wrong direction (A-F) and no-payments(G-H).

result is shown in Fig.12 ($\lambda = \lambda_2 = 40$), obviously, the result of group sparsity in Fig.12(a) is sparse, we can easily select features of B with score $\|X_{i.}\|_2^2 > 0$ as dictionary; however, in Fig.12(b), as we use Frobenius norm, nearly the score of all the features are greater than zero, which makes it hard to select the dictionary. In Fig.12(c) and Fig.12(d), we rerank the score, the result of our group sparsity model using $L_{2,1}$ norm automatically select about 250 features from 900 features as dictionary, and in contrast the result of dictionary selection model using Frobenius norm cannot work well. Fig.12 is only one of our experiments, in other cases, the results are also similar, thus we can conclude that our dictionary selection model using group sparsity, $L_{2,1}$, is effective.

7. Conclusion

In this paper, we propose a new criterion, sparse reconstruction cost (SRC), for abnormal event detection in the crowded scene. Whether a testing sample is abnormal or not is determined by its sparse reconstruction cost, through a weighted linear reconstruction of the over-completed normal bases. Our proposed dictionary selection method supports a robust estimation of the dictionary with minimal size; and with the help of the low rank constraint, it can not only deal with large scale training samples, but also require less memory cost than our previous work [7]. Thanks to the flexibility in designing the basis, our method can easily handle



(a) Coefficients of dictionary selection using $L_{2,1}$ norm

(b) Coefficients of dictionary selection using Frobenius norm

(c) Rerank the coefficients of dictionary selection using $L_{2,1}$ norm

(d) Rerank the coefficients of dictionary selection using Frobenius norm

Figure 12: We compare the dictionary selection model using $L_{2,1}$ norm with Frobenius norm. The result using group sparsity, i.e. $L_{2,1}$ norm, is sparse and effective.

both local abnormal events (LAE) and global abnormal events (GAE). By incrementally updating the dictionary, our method also supports online event detection. The experiments on three benchmark datasets show favorable results when compared with the state-of-the-art methods. In fact, our algorithm provides a general solution for outlier detection; and can also be applied to other applications, such as event/action recognition.

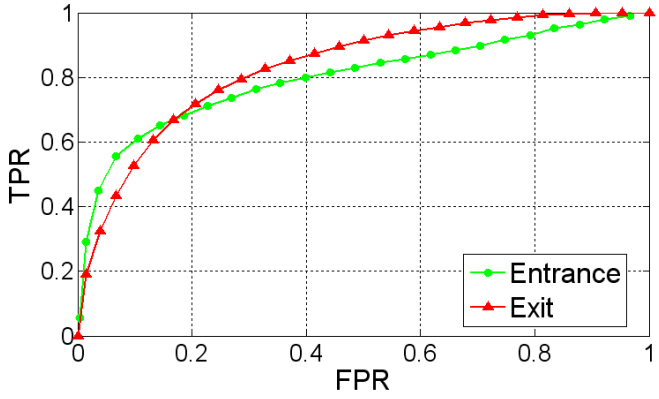


Figure 13: The frame-level ROC curve for both subway entrance and exit datasets

Appendix A. Appendix A

We prove Theorem 1 here, where the optimization problem $\min_{\mathrm{X}}: \ p_{Z,L}(\mathrm{X})$ can be equivalently written as:

$$\begin{aligned}
&\min_{\mathrm{X}}: \ f_0(\mathrm{Z}) + \langle \nabla f_0(\mathrm{Z}), \mathrm{X}-\mathrm{Z}\rangle + \frac{L}{2}\|\mathrm{X}-\mathrm{Z}\|_F^2 + \lambda\|\mathrm{X}\|_{2,1} \\
\Leftrightarrow &\min_{\mathrm{X}}: \ \frac{L}{2}\|(\mathrm{X}-\mathrm{Z}) + \frac{1}{L}\nabla f_0(\mathrm{Z}))\|_F^2 + \lambda\|\mathrm{X}\|_{2,1} \\
\Leftrightarrow &\min_{\mathrm{X}}: \ \frac{L}{2}\|\mathrm{X}-(\mathrm{Z}-\frac{1}{L}\nabla f_0(\mathrm{Z}))\|_F^2 + \lambda\|\mathrm{X}\|_{2,1} \\
\Leftrightarrow &\min_{\mathrm{X}}: \ \frac{L}{2}\|\mathrm{X}-(\mathrm{Z}-\frac{1}{L}\nabla f_0(\mathrm{Z}))\|_F^2 + \lambda\sum_{i=1}^{k}\|\mathrm{X}_{i.}\|_2
\end{aligned} \tag{A.1}$$

Since the L_2 norm is self dual, the problem above can be rewritten by introducing a dual variable $\mathrm{Y} \in \mathbb{R}^{k\times k}$:

$$\begin{aligned}
&\min_{\mathrm{X}}: \ \frac{L}{2}\|\mathrm{X}-(\mathrm{Z}-\frac{1}{L}\nabla f_0(\mathrm{Z}))\|_F^2+\lambda\sum_{i=1}^{k}\max_{\|\mathrm{Y}_{i.}\|_2\le 1}\langle \mathrm{Y}_{i.},\mathrm{X}_{i.}\rangle \\
\Leftrightarrow &\max_{\|\mathrm{Y}_{i.}\|_2\le 1}\min_{\mathrm{X}}: \ \frac{L}{2}\|\mathrm{X}-(\mathrm{Z}-\frac{1}{L}\nabla f_0(\mathrm{Z}))\|_F^2+\lambda\sum_{i=1}^{k}\langle \mathrm{Y},\mathrm{X}\rangle \\
\Leftrightarrow &\max_{\|\mathrm{Y}_{i.}\|_2\le 1}\min_{\mathrm{X}}: \ \frac{1}{2}\|\mathrm{X}-(\mathrm{Z}-\frac{1}{L}\nabla f_0(\mathrm{Z})-\frac{\lambda}{L}\mathrm{Y})\|_F^2 \\
&-\frac{1}{2}\|\mathrm{Z}-\frac{1}{L}\nabla f_0(\mathrm{Z})-\frac{\lambda}{L}\mathrm{Y}\|_F^2
\end{aligned} \tag{A.2}$$

The second equation is obtained by swapping "max" and "min". Since the function is convex with respect to X and concave with respect to Y, this swapping does not change the problem by the Von Neumann minimax theorem. Letting $\mathrm{X}=\mathrm{Z}-\frac{1}{L}\nabla f_0(\mathrm{Z})-\frac{\lambda}{L}\mathrm{Y}$, we obtain an equivalent problem from the last equation above

$$\max_{\|\mathrm{Y}_{i.}\|_2\le 1}: \ -\frac{1}{2}\|\mathrm{Z}-\frac{1}{L}\nabla f_0(\mathrm{Z})-\frac{\lambda}{L}\mathrm{Y}\|_F^2 \tag{A.3}$$

Using the same substitution as above,

$$\mathrm{Y}=-\frac{L}{\lambda}(\mathrm{X}-\mathrm{Z}+\frac{1}{L}\nabla f_0(\mathrm{Z})), \tag{A.4}$$

we change it into a problem in terms of the original variable X as

$$\min_{\|\frac{L}{\lambda}(\mathrm{X}-\mathrm{Z}+\frac{1}{L}\nabla f_0(\mathrm{Z}))_{i.}\|_2\le 1}: \ \|\mathrm{X}\|_F^2 \Leftrightarrow \sum_{i=1}^{k}\min_{\|\mathrm{X}_{i.}-(\mathrm{Z}-\frac{1}{L}\nabla f_0(\mathrm{Z}))_{i.}\|_2\le\frac{\lambda}{L}}: \ \|\mathrm{X}_{i.}\|_2^2. \tag{A.5}$$

Therefore, the optimal solution of the first problem in Eq.(A.5) is equivalent to the last problem in Eq.(A.5). Actually, each row of X can be optimized independently in the last problem. Considering each row of X respectively, we can get the closed form as

$$\arg\min_{\mathrm{X}} p_{\mathrm{Z},L}(\mathrm{X})=\mathscr{D}_{\frac{\lambda}{L}}(\mathrm{Z}-\frac{1}{L}\nabla f_0(\mathrm{Z})). \tag{A.6}$$

Appendix B. Appendix B

As Frobenius norm can be considered as a kind of L_2 norm, it can be rewritten as $\|\mathrm{X}\|_F^2=tr(\mathrm{X}^T\mathrm{X})$, where $tr(\mathrm{A})=\sum_i \mathrm{A}_{ii}$ is the trace of matrix A. Thus, we can

rewrite Eq.(23) as

$$F_s = \arg\min_{\mathrm{B}} tr((\mathrm{B}-\mathrm{BX})^T(\mathrm{B}-\mathrm{BX})) + \lambda_2\, tr(\mathrm{X}^T\mathrm{X}), \tag{B.1}$$

where $\mathrm{B} \in \mathbb{R}^{m\times k}$ and $\mathrm{X} \in \mathbb{R}^{k\times k}$. In order to solve this equation, we derivative it,

$$\frac{\partial F_s}{\partial \mathrm{X}} = 0. \tag{B.2}$$

Obviously, this is a convex optimization, and the quadratic optimization can be used to solve it. As

$$\frac{\partial tr(\mathrm{AB})}{tr(\mathrm{A})} = \frac{\partial tr(\mathrm{BA})}{tr(\mathrm{A})} = \mathrm{B} \tag{B.3}$$

We have

$$\frac{\partial tr((\mathrm{B}-\mathrm{BX})^T(\mathrm{B}-\mathrm{BX})) + \lambda_2\, tr(\mathrm{X}^T\mathrm{X})}{\partial \mathrm{X}} = 0, \tag{B.4}$$

$$\frac{\partial tr(\mathrm{B}^T\mathrm{B}-\mathrm{B}^T\mathrm{BX}-\mathrm{X}^T\mathrm{B}^T\mathrm{B}+\mathrm{X}^T\mathrm{B}^T\mathrm{BX}) + \lambda_2\, tr(\mathrm{X}^T\mathrm{X})}{\partial \mathrm{X}} = 0. \tag{B.5}$$

So, we can get

$$-2\mathrm{B}^T\mathrm{B} + 2\mathrm{B}^T\mathrm{B}X + \lambda_2 2\mathrm{X} = 0, \tag{B.6}$$

that is

$$(\mathrm{B}^T\mathrm{B} + \lambda_2\mathrm{I})\mathrm{X} = \mathrm{B}^T\mathrm{B}, \tag{B.7}$$

where $\mathrm{I} \in \mathbb{R}^{k\times k}$ is an identity matrix. Usually, $\lambda_2 > 0$, so $(\mathrm{B}^T\mathrm{B} + \lambda_2\mathrm{I})$ is a full rank matrix and has an inverse matrix, therefore we have a close-form solution of X for Eq.(23) as

$$\mathrm{X} = (\mathrm{B}^T\mathrm{B} + \lambda_2\mathrm{I})^{-1}\mathrm{B}^T\mathrm{B}. \tag{B.8}$$

[1] Adam, A., Rivlin, E., Shimshoni, I. and Reinitz, D. [2008], 'Robust real-time unusual event detection using multiple fixed-location monitors', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **30**, 555–560.

[2] Avidan, S. [2007], 'Ensemble tracking', *IEEE transactions on pattern analysis and machine intelligence* pp. 261–271.

[3] Benezeth, Y., Jodoin, P., Saligrama, V. and Rosenberger, C. [2009], Abnormal events detection based on spatio-temporal co-occurences, *in* 'CVPR'.

[4] Boiman, O. and Irani, M. [2005], Detecting irregularities in images and in video, *in* 'ICCV'.

[5] Boiman, O. and Irani, M. [2007], 'Detecting irregularities in images and in video', *International Journal of Computer Vision* **74**(1), 17–31.

[6] Cong, Y., Gong, H., Zhu, S. and Tang, Y. [2009], Flow mosaicking: Real-time pedestrian counting without scene-specific learning, *in* 'CVPR', pp. 1093–1100.

[7] Cong, Y., Yuan, J. and Liu, J. [2011], Sparse Reconstruction Cost for Abnormal Event Detection, *in* 'IEEE Conf. on Computer Vision and Pattern Recognition', pp. 3449–3456.

[8] Dalal, N. and Triggs, B. [2005], Histograms of oriented gradients for human detection, *in* 'CVPR', pp. 886–893.

[9] D.Helbing, P. [1995], 'Social force model for pedestrian dynamics', *Physical Review* **E, 51**, 4282.

[10] Ernesto L. Andrade, S. B. and Fisher, R. B. [2006], Modelling crowd scenes for event detection, *in* 'ICPR'.

[11] *http://www.svcl.ucsd.edu/projects/anomaly* [n.d.].

[12] Hu, W., Xiao, X., Fu, Z., Xie, D., Tan, T. and Maybank, S. [2006], 'A system for learning statistical motion patterns', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **28**(9), 1450–1464.

[13] Huang, K. and Aviyente, S. [2007], Sparse representation for signal classification, *in* 'NIPS'.

[14] Itti, L. and Baldi, P. [2005], A principled approach to detecting surprising events in video, *in* 'CVPR'.

[15] Jiang, F., Yuan, J., Tsaftaris, S. and Katsaggelos, A. [2011], 'Anomalous video event detection using spatiotemporal context', *Computer Vision and Image Understanding* **115**(3), 323–333.

[16] Kim, J. and Grauman, K. [2009], Observe locally, infer globally: A space-time MRF for detecting abnormal activities with incremental updates, *in* 'CVPR'.

[17] Kratz, L. and Nishino, K. [2009], Anomaly detection in extremely crowded scenes using spatio-temporal motion pattern models, *in* 'CVPR'.

[18] Lee, H., Battle, A., Raina, R. and Ng, A. [2007], 'Efficient sparse coding algorithms', *NIPS* **19**, 801.

[19] Liu, C., Freeman, W., Adelson, E. and Weiss, Y. [2008], Human-assisted motion annotation, *in* 'CVPR'.

[20] Liu, J., Musialski, P., Wonka, P. and Ye, J. [2009], Tensor completion for estimating missing values in visual data, *in* 'ICCV'.

[21] Liu, J., Wonka, P. and Ye, J. [2011], 'Sparse non-negative tensor factorization using columnwise coordinate descent', *Pattern Recognition* .

[22] Ma, Y., Derksen, H., Hong, W. and Wright, J. [2007], 'Segmentation of multivariate mixed data via lossy data coding and compression', *IEEE Trans. on Pattern Analysis and Machine Intelligence* .

[23] Mahadevan, V., Li, W., Bhalodia, V. and Vasconcelos, N. [2010], Anomaly detection in crowded scenes, *in* 'CVPR', pp. 1975–1981.

[24] Mei, X. and Ling, H. [2009], Robust visual tracking using l1 minimization, *in* 'ICCV', IEEE, pp. 1436–1443.

[25] Olshausen, B. and Field, D. [1997], 'Sparse coding with an overcomplete basis set: A strategy employed by v1?', *Vision research* **37**(23), 3311–3325.

[26] Olshausen, B. et al. [1996], 'Emergence of simple-cell receptive field properties by learning a sparse code for natural images', *Nature* **381**(6583), 607–609.

[27] Ramin Mehran, Alexis Oyama, M. S. [2009], Abnormal crowd behavior detection using social force model, *in* 'CVPR'.

[28] Saad Ali, M. S. [2008], Floor fields for tracking in high density crowd scenes, *in* 'ECCV'.

[29] Stauffer, C. and Grimson, W. [2002], Adaptive background mixture models for real-time tracking, *in* 'CVPR'.

[30] Tziakos, I., Cavallaro, A. and Xu, L. [2010], 'Event monitoring via local motion abnormality detection in non-linear subspace', *Neurocomputing* **73**, 1881–1891.

[31] *Unusual crowd activity dataset of University of Minnesota, from http://mha.cs.umn.edu/movies/crowdactivity-all.avi.* [n.d.].

[32] Wang, X., Ma, X. and Grimson, W. [2009], 'Unsupervised activity perception in crowded and complicated scenes using hierarchical Bayesian models', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **31**(3), 539–555.

[33] Wright, J., Yang, A., Ganesh, A., Sastry, S. and Ma, Y. [2008], 'Robust face recognition via sparse representation', *TPAMI* **31**(2), 210–227.

[34] Wu, S., Moore, B. and Shah, M. [2010], Chaotic invariants of Lagrangian particle trajectories for anomaly detection in crowded scenes, *in* 'CVPR'.

[35] W.Yu, A. J. [2007], 'Modeling crowd turbulence by many-particle simulation', *Phycical Review E* **76(4)**, 046105.

[36] Yuan, J., Liu, Z. and Wu, Y. [2009], Discriminative subvolume search for efficient action detection, *in* 'Proc. IEEE Conf. on Computer Vision and Pattern Recognition'.

[37] Zhang, T. [2009], 'On the consistency of feature selection using greedy least squares regression', *The Journal of Machine Learning Research* **10**, 555–568.

[38] Zhao, B., Fei-Fei, L. and Xing, E. [2011], Online detection of unusual events in videos via dynamic sparse coding, *in* 'CVPR'.

[39] Zhong, H., Shi, J. and Visontai, M. [2004], Detecting unusual activity in video, *in* 'CVPR'.